学校编码：10384　　分类号_____密级_____
学　　号：20051302340　　UDC _____

厦 门 大 学

硕 士 学 位 论 文

基于贝叶斯网络的中医医案数据挖掘

Data Mining in TCM Medical Records Based on Bayesian Network

李 长 军

指 导 教 师：李 绍 滋 教授
专 业 名 称：计 算 机 应 用
论文提交日期：2008 年　　月
论文答辩日期：2008 年　　月
学位授予日期：2008 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2008 年　月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非盈利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于
1、保密（　），在　　年解密后适用本授权书。
2、不保密（　）
（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日
导师签名：　　日期：　　年　　月　　日

摘要

中医作为中华民族的瑰宝，距今已经有几千年的发展历史。各代医家在长期的临床实践中积累了大量的经验，对这些经验的总结不但能丰富中医的理论体系，还能对中医的现代化起到巨大的促进作用。因此，对中医各种资料的总结、挖掘和利用是我国一项非常重要的任务。然而，传统的方法通过口传身受、手工整理难以做到系统化、客观化。因此，有必要用计算机技术对中医药资料进行挖掘利用，整理归纳中医药资料隐含的规律。

贝叶斯网络作为一种被广泛成功应用的数据挖掘技术，其在医疗行业也有很成功的应用。本文以慢性胃炎临床诊断病例以及近现代名老中医“内生五邪”医案为原始数据，针对贝叶斯网络应用到这些数据中碰到的问题，本文提出了数据预处理和分组降维等解决方法，具体工作如下：

（1）分析探讨了中医医案的特点，对慢性胃炎病例和“内生五邪”医案做规范化处理，针对中医医案难以量化的特点，提出症状分解的方法，为中医数据处理提供了一条解决思路。

（2）分析了慢性胃炎病例数据的症状维高，数据分布稀疏的特点，探讨了主成分分析法在此类数据的降维效果。针对慢性胃炎病例数据的特点，提出并详细描述了层次聚类和主成分分析法相结合的分组降维方法。最后分析了该分组降维方法对贝叶斯网络学习能力的提升以及存在的问题。

（3）描述了“内生五邪”医案的数据缺失问题，详细介绍了缺失数据的定义以及对贝叶斯网络学习的影响。提出了一种基于模拟退火和 BC 算法的改进的 SEM 算法（E-SEM）。最后把 E-SEM 算法应用到“内生五邪”医案的“五邪”识别并对算法性能做了分析。

关键词：贝叶斯网络；中医药数据挖掘；缺失数据

Abstract

As a kind of treasures in our country, Traditional Chinese Medicine (TCM) has a history of thousands of years. Herbalist doctors in each generation have accumulated a wealth of experience in long-term clinical practice. The summary of precious experience can not only enrich the TCM theory system, but also can promote the modernization of TCM. Therefore, the summary, exploitation and implementation of TCM data are a very important task in our country. However, it is difficult to be systematic and objective when using traditional methods such as masters' impartation and manual processing to process TCM data. So, it is necessary to use computer technology to exploit TCM data and find laws behind these data.

As a widely and successfully applied technology of data mining, Bayesian Network also has successful application in the medical field. In this thesis, original data for research came from clinical cases of chronic gastritis and medical records of modern famous herbalist doctors. According to the problems when Bayesian Network dealt with these data, this thesis gave methods of data preprocessing and grouping dimension-reduction. The main work is as follows:

(1) The characteristic and normalization of TCM medical records are discussed in this thesis. As to the problem that TCM medical records are hard to be quantified, this thesis gave a quantification method which use symptom decomposition. The proposed method gave a way to process the TCM data.

(2) The feature of high symptom dimension and sparse data in chronic gastritis cases was analyzed. Because of this feature, the symptom dimension of chronic gastritis is hard to be reduced. To solve this problem, this thesis proposed and described a grouping dimension-reduction method which combines hierarchy clustering with principal component analysis (PCA). Finally, the proposed method is applied to boost the learning ability of Bayesian Network. The experiment results and some open problems were also discussed.

(3) This thesis described data-missing problems in medical records of "internal generation of five pathogenic factors". Then, this thesis gave a comprehensive

discussion of definition of missing data and impact which missing data caused on Bayesian Network learning. Following the introduction of many processing method of missing data, this thesis gave an improved SEM algorithm (E-SEM) based on Simulating Annealing and BC algorithm. Finally, the improved algorithm is used to train Bayesian Network and recognize the “five pathogenic factors” in the medical records. The performance of E-SEM algorithm was also discussed.

Key words: Bayesian Network; Data Mining in TCM Data; Missing Data

目录

Contents

第一章 绪论

1.1 研究背景与意义

中医作为中华民族的瑰宝，距今已经有几千年的发展历史。通过各代医家的继承和发展，中医逐渐形成了自己独特的理论体系，也积累了大量的医案、专著和药方等资料，为中医的继续发展提供了丰富而宝贵的信息。

然而，由于中医偏重主观分析和个体经验，中医药资料大都具有主观性、不规范性以及模糊性等特点，使得中医药资料无法得到充分的挖掘和利用。造成这种问题的根源在于中医“辨证施治”的思想。“辨证施治”就是对四诊（望、闻、问、切）所得资料进行综合分析，概括、判断出反映疾病本质特征的证候，从而对病人进行具体施治的思想和方法，其本质在于强调病人的个体差异以及致病因素的综合分析。“辨证施治”为中医诊治提供了有效的思想方法，但也从某种程度上阻碍了中医的推广。中医对症状和证候的判断带有主观性且症状和证候难以量化，这对于崇尚量化和可重复性实验的现代医学来说是难以理解的。由于此，很多人质疑中医的科学性，中医也出现了日渐衰败的迹象。

中医在几千年的发展中，能够生生不息，代代相传，应该说具有其存在的原因和必然性。同时，中医通过时间检验也说明其具有一定的科学意义，只是不能用现代科学进行完全解释。因此，利用现代科学技术，特别是数据挖掘方法对中医药数据进行整理、归纳以及知识抽取，从而对中医进行科学的分析和阐释，这不但是中医发展和现代化的需要，也是历史赋予我们的责任。

数据挖掘是当今活跃的具有广阔应用前景的信息技术研究领域，是人工智能、统计学、机器学习、认知科学、并行计算和数据可视化等多领域相互交叉的研究方向[1]。作为一种有效的知识发现方法，数据挖掘在电信、金融、医疗等各个行业都有着成功的应用。对于中医领域来说，由于中医药资料的不规范性和模糊性等特点，传统的单一的方法无法对中医药资料进行有效的分析。在数据挖掘技术被广泛而成功的应用的背景下，利用数据挖掘技术对中医药资料进行挖掘和研究就成为了一种很好的发展中医的思路。目前，已经有一些研究

人员在这方面进行了广泛有益的探索[2-4]。

贝叶斯网络作为数据挖掘领域的一种方法，由于其具有知识表达自然直观，推理灵活以及能方便处理缺失数据等优点，一直以来都是数据挖掘领域广泛采用的方法[5]。对于中医来说，利用贝叶斯网络对中医药数据进行数据挖掘有以下两个方面的优势：1.利用贝叶斯网络建模，所得模型易于理解，这对中医来说至关重要。2. 贝叶斯网络能够方便地处理缺失数据，而中医药资料一般都存在数据缺失的情况。考虑到贝叶斯网的诸多优点，贝叶斯网络可以作为一种很好的中医药数据挖掘的方法。所以，使用贝叶斯网络对中医药数据进行挖掘，对中医推广和现代化都具有重要的现实意义。

本论文以中医慢性胃炎和“内生五邪”医案证型识别为背景，利用贝叶斯网络方法对整理好的医案数据库进行中医规律的挖掘和分析，以期能够为贝叶斯网络的算法研究以及在中医的应用提供一些思路和借鉴。

1.2 数据挖掘技术在中医药领域的发展现状和存在的问题

1.2.1 数据挖掘技术在中医药领域的应用

随着中医规律研究的不断深入，数据挖掘技术逐步被中医研究人员所接受，并逐步将其应用到相关研究中，同时获得了一些可喜的成果。

瞿海斌等利用决策树从 290 例血癖证病例中自动提取相应的诊断规则，得到决策树分类模型，并归纳出 5 条血癖证的诊断规则，对 194 例血癖证病例测试结果为：阳性检测正确率、阴性检测正确率和检测正确率分别达到 97.67%、99.07%和 98.45%。实验结果表明决策树能自动从中医病例中归纳诊断规则[6]。

浙江大学 Wang Huiyan 等人利用贝叶斯网络对脉象仪采集到的脉象信号进行脉象识别，实现了脉象的自动诊断[2]。该论文的方法为脉象的客观检测提供了很好的思路。浙江大学王学伟等人利用 MCMC 采样方法改进贝叶斯网络的学习算法，生成的贝叶斯网络用于肺病的中医证型分类。试验结果表明，该算法能够很好地应用到中医的分类问题中[3]。哈尔滨工业大学的 Pang Bo 等人利用贝叶斯网络学习算法对从舌象图片提取出来的信息进行处理，生成的贝叶斯网络能够较准确的诊断脑梗塞、心脏病、上呼吸道感染等 13 种常见病，其中对胰